Archiving in the Networked World: Preserving Plagiarized Works

by Michael Seadle
Editor, *Library Hi Tech*
Professor and Director, Berlin School of Library and Information Science

ABSTRACT

Type: research article

Purpose: Plagiarism has become a salient issue for universities and thus for university libraries in recent years. This article discusses three interrelated aspects of preserving plagiarized works: collection development issues, copyright problems, and technological requirements. Too often these three are handled separately even though in fact each has an influence on the other.

Methodology: The article looks first at the ingest process (called the Submission Information Package or SIP, then at storage management in the archive (the AIP or Archival Information Package), and finally at the retrieval process (the DIP or Distribution Information Package).

Findings: The chief argument of this article is that works of plagiarism and the evidence exposing them are complex objects, technically, legally and culturally. Merely treating them like any other work needing preservation runs the risk of encountering problems on one of those three fronts.

Implications: This is a problem, since currently many public preservation strategies focus on ingesting large amounts of self-contained content that resembles print on paper, rather than on online works that need special handling. Archival systems also often deliberately ignore the cultural issues that affect future usability.

Introduction

Plagiarism has become a salient issue for universities and thus for university libraries in recent years. The problem has especially been in the news in Germany, where the former Defense Minister, Karl Theodor Freiherr zu Guttenberg, lost both his doctorate and his job when evidence accumulated that much of his 2009 dissertation “Verfassung und Verfassungsvertrag: Konstitutionelle Entwicklungsstufen in den USA und der EU” consisted of text copied from other sources. While it is easier to plagiarize texts using cut-and-paste in

the networked world, it is also easier to discover plagiarism using both commercial tools and public search engines like Google or Bing. Guttenberg is by no means the only example of a public figure whose position was lost as a consequence of accusations of plagiarism, but his case makes a good and very public example to use to consider what plagiarism means for long term digital archiving. Note that this article makes no judgment itself about whether Guttenberg plagiarized his thesis – the accusation comes from media reports and other publicly available sources.

This article discusses three interrelated aspects of preserving plagiarized works: collection development issues, copyright problems, and technological requirements. Too often these three are handled separately even though in fact each has an influence on the other. The discussion will integrate these three themes using the structure of the OAIS (Open Archival Information System) reference model (CCSDS, 2002). The article will look first at the ingest process (called the Submission Information Package or SIP, then at storage management in the archive (the AIP or Archival Information Package), and finally at the retrieval process (the DIP or Distribution Information Package).

Ingest

The key decision at ingest is whether a work should be preserved for posterity. As a general rule digital archiving tries to preserve scholarly works and works of significant social importance. A work of plagiarism is not generally acceptable as scholarship, but in the case of Guttenberg, the dissertation has social significance because of his high government post and because of the public controversy that the plagiarism accusations stirred. Prior to the controversy Guttenberg had a particularly strong reputation for honesty. Von Darnstädt et al. (2011) reported in *Spiegel*:

> *Er ist der Politiker, der als besonders ehrlich, aufrichtig und authentisch galt. Im "Ehrlichkeitsranking" des Wissenschaftlers Horst Opaschowski belegte er von allen aktiven Politikern den besten Rang. Emnid hatte im September 2010 ermittelt, dass 67 Prozent der Deutschen Guttenberg für glaubwürdig hielten, 69 Prozent für geradlinig. Seit längerem war er der beliebteste Politiker in Deutschland, 79 Prozent wollten ihn in einer wichtigen Rolle sehen*.
>
> He is a politician, who counted as especially honest, upright and authentic. In the "honesty-ranking" of the scholar Horst Opaschowski he counted as the best of all the active politicians. Emnid [a German survey research firm] reported in September 2010 that 67% of Germans considered Guttenberg trustworthy, and 69% straightforward. He was long the most beloved politician in Germany; 79% wanted to see him in a more important job. [My translation}

Had Guttenberg been less well trusted, and less well regarded as a strong contender for the office of Chancellor, the accusation of plagiarism would have made fewer headlines. In Guttenberg's case the scandal became a political event. That alone might be reason enough for libraries to want to preserve the original document.

Guttenberg's alleged plagiarism also had an effect on German universities, which suddenly began receiving questions and accusations about other potential plagiarism cases. A number of universities established or reestablished commissions on academic integrity to investigate and decide about these accusations (full disclosure: I am a member of such a commission at my home institution). Some universities also began to discuss whether to change the review process for dissertations in order to require a member of the committee to come from the outside. The intention was that such a person would be less influenced by personal ties and perhaps therefore more apt to spot problems. Whether such measures would discourage plagiarism is not the point. The point is that Guttenberg's plagiarism triggered changes and discussions within the academic community, which is also a reason for preserving a document.

Assuming then that preserving at least this work of plagiarism as an historical document is reasonable, the questions arise: 1) what institution ought to initiate the preservation and 2) what exactly should be preserved.

Two possible institutions come quickly to mind. One is the University of Bayreuth, where Guttenberg received his doctorate and which ultimately took it away from him. For Bayreuth the dissertation has special importance because of the embarrassment it caused. For that reason the university could also prefer not to keep it around. It has kept it, however. As of September 2011 the dissertation remains in the library according to its online catalog and it may be loaned out. The second with a potential preservation interest is the German National Library (Deutsche Nationalbibliothek or DNB) because of the political effect of the plagiarism accusations. The DNB also continues to have a record in its catalog. This suggests that either or both libraries might well choose to preserve it. Indeed there is some indication that demand for it has increased – a single used copy on Amazon is for sale for 380 €, over four times the original price of 88€. A third possibility for preservation could be the publisher, but Duncker & Humblot have withdrawn the work, which makes the firm an unlikely preservation candidate.

The choice of what exactly to preserve seems simple at one level: the dissertation itself as it was approved at Bayreuth and as it was published in 2009. Merely preserving the work itself does not, however, preserve either the social, political or cultural context in which the accusations of plagiarism occurred, or the evidence of copying. For the evidence, a future reader would need a website called GutenPlag Wiki [1] that documents instances of plagiarism in the dissertation: "1218 plagiarized fragments out of 135 sources on 371 out of 393 pages (94.4%) in 10421 lines (63.8%)" [my translation]. (Note: "Plag" in GutenPlag comes from "Plagiat", the German word for plagiarism.) GutenPlag Wiki is complex to

preserve because it is dynamic and could theoretically still change, though the last new change took place in April 2011. GutenPlag Wiki also has its own copyright protection.

Under German law the German National Library as a depository library has a clear right to preserve Guttenberg's original dissertation. Its right to preserve GutenPlag Wiki is less straightforward, though the founders of GutenPlag Wiki may not object. The technical issues in preserving a PDF of Guttenberg's thesis are unproblematic, but preserving a dynamic site like GutenPlag Wiki involves multiple programs and interactive relationships (for example, the detailed listing of where plagiarism occurred in the text and what the source was is not actually on GutenPlag Wiki, but at a website with the name "gut.greasingwheels.org"). Some digital preservation systems, notably LOCKSS (Lots of Copies Keep Stuff Safe from Stanford University) can routinely preserve complex information objects, but many archives have been designed primarily for simple file-based data such as a PDF. The site could be rendered as a PDF to preserve it on a set date and time, but that would lose some of its dynamic character, such as the image of the pages of the dissertation that slowly fill with colored indicators of the plagiarized passages each time the web page loads.

Once all of the copyright and technical problems are resolved, the further question arises whether to package Guttenberg's dissertation in its original form and GutenPlag Wiki together in an archival unit (SIP) in order to preserve the contextual relationship between them, or whether these two different types of works, one a traditional print product, the other a website, should be packaged in archival units with other works of similar format, similar contents, or similar publication dates. This question has no simple answer, since most archives have well-established guidelines for how to construct their SIP packages. Constructing a SIP that ignores these rules could potentially also make discovery technically more complex. Whether packaging the dissertation and the evidence together makes sense is an intellectual and technical choice with long term consequences. Maintaining the contextual relationship is certainly possible in separate archival units, but the link would arguably be more fragile.

Archival Management

The issues for a plagiarized work within an archival system are not fundamentally different than for other works, but a few are worth special consideration, especially those involving authenticity and the ownership of copyrights.

Authenticity is a problem from the outset for a plagiarized work, because it is in some sense not entirely authentic to begin with, but rather a patchwork of original and copied elements. The Guttenberg case illustrates this well. Nonetheless an archival system should be able to maintain the authenticity of Guttenberg's dissertation in the form that he published it when getting his doctorate, if a PDF of the document were held in a system like LOCKSS (Lots of Copies Keeps Stuff Safe) that keeps multiple copies and checks them regularly for integrity to make sure that no change has occurred. One part of authenticity is evidence that the

document is unchanged, since a document with changes is clearly not authentic. Another part is the provenance record of copies made and checked over time, which needs also to be kept in some independent location (or rather locations) over time. This is authenticity at its most basic.

A work of plagiarism that has political importance, such as Guttenberg's dissertation, offers a potential temptation at some future date to those who might want to rescue the author's reputation (or to make it worse) by tampering with the text. A group could, for example, create a version of the dissertation without the plagiarism (or with more plagiarism), archive it, and then claim that it was the true and authentic version. Such a scenario may seem unlikely at this time, but it is precisely because of these kinds of unlikely possibilities that digital preservation needs mechanisms to ensure authenticity. A paper copy of the dissertation may well not exist in 100 years. Trusted institutions could play a role, but trust itself is a political choice and any institution can be corrupted. In terms of future evidence the number of identical copies is likely to play a role, as well as the record of provenance and the degree to which the storage systems are reasonably robust against human intervention (post submission) and against external attack.

In the case of Guttenberg's dissertation the GutenPlag Wiki provides an additional authenticity check, since it has a full record of where plagiarism occurred on each page of the original dissertation. GutenPlag Wiki is more than just an additional copy, because it draws the connection between passages in the dissertation and the original published text in other sources. Yet this connection works reliably for authenticity checking only if the other works also exist in an archive and if those works can also make a reasonable claim to authenticity. And it works well only if the linkages in GutenPlag Wiki continue to function. Relatively few archiving systems have made a serious effort to ensure that IP-based references continue to have meaning over time.

Maintaining the functionality of IP-based references over time is a problem as serious as format migration. A work whose normal functioning depends in part on links is not really fully usable when they fail. This is one of the chief differences between digitized versions of paper-based documents and true Internet content. Maintaining this functionality has multiple layers. One level of functionality is within the set of services that provide content and formatting for the basic screen images (HTML, CSS, Javascript, etc.). Another level is within the organizational or institutional context to enable links to higher level or lower level pages. GutenPlag Wiki has links that go outside of its own organization. These links will work in 100 years only if the domain name resolvers are also archived, and if they are archived in such a way that those numeric IP addresses whose URL names have changed still resolve to the right address for the archived content. This problem is by no means unique to works of plagiarism, but it certainly matters when trying to preserve GutenPlag Wiki with 135 external references.

The copyright ownership issue for Guttenberg's dissertation is not clear, since the accusation is that he did not author large portions of the work. The issue could be resolved during ingest, but there is a good chance that it will be left for the future, especially in a dark archive. That is reasonable enough, if the information needed for resolving the copyright ownership issues is also archived. It could be argued that each author whose work Guttenberg copied has a copyright on that particular portion of the text and that the text should not be available for public domain use until the rights of all authors have expired..

An attempt to preserve the information necessary to resolve all possible copyright claims means having a source that supplies the death dates of all possible authors. It also ideally means maintaining an access rule base that keeps up-to-date about changes to the applicable laws. If the term of protection were to be extended beyond the life of the author plus 70 years, then information about the various authors' heirs and about possible contractual arrangements could also be needed in order to make the content legally and safely available within a century.

In Europe the copyright laws tend to be retroactive, so that the current law applies to works created before the law's enactment. This has not generally been true in the US, where some works still have protection under the old 1909 copyright law. That means that some of the US works within the GutenPlag Wiki could potentially, in 100 years, have a different length or type of protection than is currently the case. It is also politically possible that the US (or other countries) choose to interpret the international copyright agreements embodied in the Berne Convention differently or to ignore them altogether. US courts have already shown some tendency toward this. The point is simply that archival systems that contains works like GutenPlag Wiki with complex international linkages should not be content with merely storing the content, if they expect to allow access in a timeframe that is under 100 years. Works of plagiarism are not unique here. They merely highlight the problem.

Retrieval

Many archival systems treat retrieval as a mechanical process: someone requests a work and the archive delivers it. This works reasonably for straightforward, self-contained scholarly works, but in the case of Guttenberg's dissertation, purely mechanical access might fail to include social and cultural elements.

One social issue that is already in the news is whether students should be allowed to cite from Guttenberg's dissertation. The University of Osnabrück was among the first German institutions to include a warning in Guttenberg's dissertation:

> *In das Werk habe das Institut für europäische Rechtswissenschaft eine Pressemitteilung der Universität Bayreuth geklebt, dass die Hochschule die an Guttenberg verliehene Doktorwürde wieder zurückgenommen habe, meldet die „Neue Osnabrücker Zeitung". Damit sollen Studenten davor gewarnt werden, aus der*

> *Arbeit zu zitieren. Das Buch kann an der Osnabrücker Universität nur noch im Lesesaal gelesen und nicht mehr ausgeliehen werden. Der Verlag hatte das Buch zwischenzeitlich eingezogen.* (Die Welt, 2011)
>
> *The Institute of European Law pasted a press clipping from the University of Bayreuth into the dissertation as a warning that the school that gave Guttenberg his doctorate has withdrawn it. The students are thus warned against citations from the work. The book can now only be read in the university library's reading room. The publisher has in the meantime withdrawn the book.* [My translation]

Der Spiegel (2011) reported a similar action at the University of Münster. Both universities were concerned that students not treat Guttenberg's work as legitimate. "Nutzer sollten aber wissen, dass die Arbeit nicht mehr als Dissertation gelte. ["Users should know that the work no longer counts as a dissertation."][My translation]".

A purely mechanical access to Guttenberg's dissertation probably would not include the press clipping about the action that the University of Bayreuth took. An archive could, in effect, paste such a clipping into the work, but that could be seen as affecting its integrity and authenticity, since the original was published without the clipping. A warning could be included in the metadata about the dissertation. The content of GutenPlag Wiki could also be delivered along with the dissertation itself. These are some of the decisions that an archive needs to make with works of plagiarism.

The universities' concern to warn students not to treat a work of plagiarism as legitimate illustrates the reaction of many contemporary academics. If Guttenberg had merely defied the copyright laws, some people might well have regarded him as a hero in defense of open access, but the fact that he took the expressions and ideas from other authors without giving them due credit goes strongly against the common culture. This seems self-evident today to most European and American academics, but the same view of copying other author's works to use in one's one is not universal.

In 2010 *Library Hi Tech* had 16 submissions where the plagiarism checking service iThenticate [2] found 20% or more overlap with other works and 10 more that had over 15% or more overlap. The high was 47%. In 2011 iThenticate found 14 submissions with 20% overlap or more, and 6 with 15% or more. The high was 75%. Not all of these submissions represented outright plagiarism. A few were cases where iThenticate misinterpreted the copying. In number of other cases the authors listed the original sources appropriately, but apparently did not understand conventions about how to indicate which passages they had copied verbatim. In still other cases the copying may have represented attempts to avoid grammatical or word-choice problems in English by taking passages from native speakers. While none of these submissions quite reached the level of copying documented in GutenPlag Wiki, the figures from iThenticate indicate that some writers in even the library

and information science community do not fully understand what plagiarism is, or why it is a problem.

Expectations across cultures are also different. Many more of the problematic *Library Hi Tech* came from Asia, where the regard for and interpretation of intellectual property appears to be different than in the US or Germany. Expectations change across time too. The US of the nineteenth century was as notorious for copyright infringement as it is now aggressive in defense of copyright – attitudes toward plagiarism could change as well. The point is that archives may wish to consider some form of digital cultural migration [3] when dealing with works of plagiarism, so that the social meaning of the plagiarism is clear to future readers who may (or may not) regard it differently.

The mechanisms for an archive to provide access to a work of plagiarism may encounter problems too. Access restrictions usually are due to copyright considerations and with works of plagiarism the copyright rules are especially complex because of the number of authors involved. Other restrictions could have to do with institutional or legal bans to protect students from works of plagiarism, or because of laws protecting the longer term reputation of the person accused of plagiarism. These restrictions are all hypothetical -- it is impossible to know what the future will bring,

The mechanics of providing access to content with multiple parts and extensive depends in part on how the archival units were packaged. If all the relevant content, in this case the PDF of Guttenberg's original dissertation, the core contents of GutenPlag Wiki and the sources of the copied texts, were in one archival unit, then all elements could simultaneously become available. If they were stored separately, then each relevant archival unit would have to be found and opened as needed. If the contents of the links were no longer available under the original URL addresses or if the whole addressing structure of Internet Protocol were to change (which is possible), then the archiving system may in effect need to provide emulation for some aspects of today's World Wide Web. The actual complexity depends on how effectively persistent identifiers really work and how well any new addressing system manages to accommodate legacy elements from the old.

Conclusion

This article has discussed special handling and special considerations for archiving works of plagiarism, but perhaps a key question is whether any strong evidence exists that ordinary archival treatment of works of plagiarism would result in serious problems. There is none, because the process of systematic long term digital archiving is scarcely more than a decade old. That is too short a time for testing most of these issues, especially those involving possible cultural miscomprehensions. It is too short even for significant amounts of published information to have vanished from the Internet. Perhaps nothing ever will, but the library and information science community is skeptical and some discussion about potential problems could help to avoid unsatisfactory results.

The chief argument of this article is that works of plagiarism and the evidence exposing them are complex objects, technically, legally and culturally. Merely treating them like any other work needing preservation runs the risk of encountering problems on one of those three fronts. Works of plagiarism are not unique in being complex and are becoming less unique as more publication becomes truly online and interactive. This is a problem, since currently many public preservation strategies focus on ingesting large amounts of self-contained content that resembles print on paper, rather than on online works that need special handling. Archival systems also often deliberately ignore the cultural issues that affect future usability. In nothing else, works of plagiarism are an indication of how current preservation strategies could fail to serve long term needs.

Notes

[1] http://de.guttenplag.wikia.com/wiki/GuttenPlag_Wiki

[2] http://www.ithenticate.com/

[3] http://digitalplusresearch.blogspot.com/2011/07/ice-forum-and-bloomsbury-conference.html

References

CCSDS (Consultative Committee for Space Data Systems) (2002), *Reference Model for an Open Archival Information System (OAIS)*, Washington DC: National Aeronautics and Space Administration. Available (September 2011): http://ddp.nist.gov/refs/oais.pdf

Die Welt Online, (9 March 2011), "Uni Osnabrück warnt vor Guttenbergs Dissertation". Available (September 2011): http://www.welt.de/politik/deutschland/article12744712/Uni-Osnabrueck-warnt-vor-Guttenbergs-Dissertation.html

Guttenberg, Karl-Theodor Freiherr zu (2009), *Verfassung und Verfassungsvertrag: Konstitutionelle Entwicklungsstufen in den USA und der EU*, Berlin: Dunker & Humblot.

Von Darnstädt, Thomas et al. (21 February 2011), „Doctor der Reserve", Der Spiegel. Available (September 2011): http://www.spiegel.de/spiegel/print/d-77108483.html